Machine Learning: Foundations **Fall Semester, 2010**

Lecture 8: December 23,2012

Lecturer: Yishay Mansour *Scribes: Liav Teichner, Idan Bassukevitz* [1]

8.1 Motivation

We wish to consider the following question: How many random examples does a learning algorithm need to sample, before it has sufficient information to learn an unknown target concept chosen from the concept class C?

In lecture 3, we presented the PAC learning model and proved some lower bounds on the number of examples required for PAC learning in several hypothesis classes. In the previous lecture, we introduced the definition of the VC-dimension, which gives us a measure of the complexity for both finite and infinite concept classes. We explored a number of examples, and showed how the VC-dimension of a hypothesis class C can be used to derive the lower bound on the number of examles required.

In this lecture we will see how the VC-dimension enables us to compute an upper bound on the number of examples required. Obviously, proving an upper bound also indicates that a learning algorithm using that bound is valid and possible.

8.2 The VC-Dimension - Review

Assume C is a concept class defined over instance space X. We can associate a concept c over X with a set (all the examples in X on which c returns a positive classification).

Let $S \subseteq X$ be a subset of X. We can define the projection of C over the subset S as follows:

Definition 8.1 *For each concept class C over X and for any subset $S \subseteq X$:*

$$\Pi_C(S) = \{c \cap S | c \in C\}.$$

Equivalently, if $S = \{x_1, \ldots, x_m\}$ then we can think of $\Pi_C(S)$ as the set of vectors and $\Pi_C(S)$ is defined by:

$$\Pi_C(S) = \{< c(x_1), \ldots, c(x_m) > | c \in C\}.$$

[1]Based in part on the scribe notes written by Shai Vardi & Tamar Lavee which in turn are based on scribe notes by Elhanan Borenstein, Orit Kliper & Ofer Pasternak which in turn are based on scribe notes by Vladimir Goldner, Yuval E. Sapir & Assaf K. Paznerin

In effect we are reducing the concept class C into the concept class $C|S$, where $S = \{x_1, ..., x_m\}$. Clearly, the concept class $C|S$ is finite with at most 2^m different concepts (as there are at most 2^m different vectors of size m) , thus:

$$|\Pi_C(S)| \leq 2^m.$$

Definition 8.2 *A concept class* C shatters S *if* $|\Pi_C(S)| = 2^m$*, or in other words a concept class shatters a set of inputs if every possible function on the input appears in* $\Pi_C(S)$*.*

Definition 8.3 VC-dim(C) (Vapnik-Chervonenkis dimension) *of* C *is the maximum size of a set* S *that is shattered by* C*:*

$$VCdim(C) = max\{d : \exists S : |S| = d \ \ and \ \ |\Pi_C(S)| = 2^d\}.$$

If C *shatters sets of arbitrarily large size (i.e such a maximum as above does not exist) we define* $VCdim(C)$ *to be infinity.*

From the bound on VC-dim above ($|\Pi_C(S)| \leq 2^m$), we can also derive that for a finite class:

$$VCdim(C) \leq \log |C|.$$

8.3 Sample Size Upper Bound

In the previous lecture we have derived a lower bound on the sample size using the VC-dimension - if $d = VCdim(C)$ then:

$$m(d, \epsilon, \delta = \frac{1}{2}) \geq \ \Omega(\frac{d}{\epsilon}).$$

We will now turn to the important application of the VC-dimension - deriving an upper bound on the sample size.

First we will show a wrong proof for an upper bound: If S is sampled than the numeber of hypotheses is $C_S = \Pi_C(S)$. C_S is finite and therefore

$$m \geq \frac{1}{\epsilon} log \frac{|Pi_C(S)|}{\delta}$$

This proof is obviously wrong (why?). We will now "fix" the proof:

Definition 8.4 *Given a target concept* c^**, the* ϵ*-bad concepts is the group of all the concepts that have an error larger than* ϵ*. Formally:*

$$B_\epsilon(c^*) = \{h \in H | error(h, c*) > \epsilon\}.$$

We will show that if we have a large enough number of samples, then none of these concepts will be consistent with the samples.

Definition 8.5 *A set of points S, is an ϵ-net for c^* with respect to a distribution D, if for each concept $h \in B_\epsilon(c^*)$ there exists a point $x \in S$ such that $h(x) \neq c^*(x)$.*

The important property of ϵ-nets is that if the sample S, sampled by a learning algorithm, forms an ϵ-net for c^*, and the learning algorithm outputs a hypothesis $h \in C$ that is consistent with S, then this hypothesis must have error less than ϵ. Thus if we can bound the probability that the random sample S fails to form an ϵ-net for c, then we have bounded the probability that a hypothesis consistent with S has error greater than ϵ.

For this discussion we will use a sample set S that is made up of two sample sets, S_1 and S_2, each is of size m, and each sampled independently according to D.

We define A as the event that S_1 is not an ϵ-net. We want to bound the probability of this event, as it bounds the probability of failure.
Assume A holds. Then there are concepts in $B_\epsilon(c)$ which are consistent with S_1. Let h be an ϵ-bad hypothesis consistent with S_1.
Now, we look at the additional sample S_2 of m points. The expectation of the error of h is at least ϵ, thus, since the median of a Binomial distibution is its expectation, with a probability of $\frac{1}{2}$, h will have more than $\frac{\epsilon m}{2}$ errors on S_2.

Define B as the event that there exists a function $h \in B_\epsilon(c)$ such that h is consistent with S_1 and has $\frac{\epsilon m}{2}$ errors on S_2. Thus:

$$Pr[B|A] \geq \frac{1}{2},$$

and by Bayes rule we get:

$$Pr[B] = Pr[B|A] \cdot Pr[A],$$

from which follows:

$$2 \cdot Pr[B] \geq Pr[A].$$

We can thus first find a bound on the probability of B, and this will apply a bound on the probability of A. The main advantage is that the event B is defined on the finite set of points $S_1 \cup S_2$.
Define F as the projection of C to $S_1 \cup S_2$. Formally:

$$F = \Pi_C(S_1 \cup S_2).$$

Later we will bound the size of F (i.e. $|F|$).

We will define the set of errors of h as follows:

$$ER(h) = \{x : x \in S_1 \cup S_2 \ \ and \ \ c(x) \neq h(x)\}.$$

We assumed that $ER(h)$ has at least $\frac{\epsilon m}{2}$ elements because in S_2 there are at least $\frac{\epsilon m}{2}$ elements from $ER(h)$. That is, $|ER(h)| \geq \frac{\epsilon m}{2}$.
Now the events can be formulated as follows:

$$\text{Event } A\text{: } ER(h) \cap S_1 = \emptyset,$$

and

$$\text{Event } B\text{: } ER(h) \cap S_1 = \emptyset \wedge ER(h) \cap S_2 = ER(h),$$

where $h \in B_\epsilon(c)$.
We wish to analyze the probability that h stays consistent with S_1 and that S_2 has at least $\frac{\epsilon m}{2}$ errors. Since we chose both S_1 and S_2 from the i.i.d. distribution D, we can build the distribution on S_1 and S_2 as follows: We sample $2m$ points $S_1 \cup S_2$ and divide the sample randomly, between S_1 and S_2. This is exactly the same distribution, because any ordering of the $2m$ elements, separated into 2 sets randomly, is the same as sampling S_1 and then S_2 (due to the i.i.d property).

Our problem is now reduced to the following simple combinatorial experiment: we have $2m$ balls (the set $S = S_1 \cup S_2$), each colored black or white, with exactly $l \geq \frac{\epsilon m}{2}$ black balls (these are the points of S that h fails on them). We divide these balls randomly into two sets of equal sizes S_1 and S_2, and we are interested in bounding the probability that all the black balls fall in S_2.

We now calculate the number of possible divisions. The number of ways we can choose l elements from $2m$ elements is:

$$\left(\begin{array}{c} 2m \\ l \end{array} \right).$$

Among them, the number of divisions in which all the black balls fall into S_2 is:

$$\left(\begin{array}{c} m \\ l \end{array} \right).$$

Thus, the probability that all of the black balls are in S_2 is exactly

$$\frac{\left(\begin{array}{c} m \\ l \end{array} \right)}{\left(\begin{array}{c} 2m \\ l \end{array} \right)} = \prod_{i=0}^{l-1} \frac{m-i}{2m-i} \leq \frac{1}{2^l}.$$

The last inequality is an approximation, assuming that each black ball can fall into S_2 with probability $\frac{1}{2}$, thus the probability that all the black balls will fall into S_2 is $\frac{1}{2^l}$. This is of course not accurate, as the black balls' probabilities are not independent (we must have exactly m balls in each group), but this is an upper bound and a very accurate approximation.

We can now bound the probabilities of A and B:

$$Pr[B] \leq |F| \cdot 2^{-l} \leq |F| \cdot 2^{-\epsilon m/2},$$

hence:

$$Pr[A] \leq 2Pr[B] \leq 2|F| \cdot 2^{-\epsilon m/2}.$$

Thus, in order for our confidence level (δ) to satisfy our goal, we will require:

$$2|F|2^{-\epsilon m/2} \leq \delta,$$

and we get that the sample size should be:

$$m = O(\frac{1}{\epsilon} \log \frac{1}{\delta} + \frac{1}{\epsilon} \log |F|).$$

The only issue we still have to resolve is a bound on the size of F. As we recall, F is a projection of c on a set with $2m$ elements. We will show that:

$$|F| = |\Pi_C(S_1 \cup S_2)| \leq (2m)^d.$$

We will define a recursion which we will later prove that it bounds the number of concept in the projection.

Definition 8.6 *Define the function J as follows:*

$$J(m, d) = J(m-1, d) + J(m-1, d-1),$$

with the initial conditions:

$$J(m, 0) = 1,$$
$$J(0, d) = 1.$$

Solving this recursion (not detailed here) gives:

$$J(m, d) = \sum_{i=0}^{d} \left(\begin{array}{c} m \\ i \end{array} \right).$$

We will use this function to bound $\Pi_c(S)$ which will yield a bound on $|F|$.

Claim 8.7 *Let $VCdim(C) = d$ and $|S| = m$, then*

$$\Pi_c(S) \leq J(m, d).$$

Proof: The proof is by induction on both d and m. For the base of the induction, the claim is easily established when $d = 0$ and m is arbitrary, and when $m = 0$ and d is arbitrary. We assume for induction that for all m', d' such that $d' + m' \leq d + m$, we have $\Pi_C(S) \leq J(m, d)$. We now show that this inductive assumption establishes the induction hypothesis for d and m.
Let $S = \{x_1, \ldots, x_m\}$ be a set of m different points and let C_S be the projection of the concept class C on S. Namely,

$$\Pi_C(S) = C_S = \{c \cap S | c \in C\}.$$

We will show that for every $S : |C_S| \leq J(m, d)$.
We define a new set T which is the set S after extracting the last point:

$$T = \{x_1, \ldots, x_{m-1}\} = S - \{x_m\} \quad , \quad |T| = m - 1$$

Define C_* as all the assignments over T which can be completed either by $x_m = 0$ or by $x_m = 1$. Then $|C_*|$ counts the number of pairs of sets in $\Pi_C(S)$ that are collapsed to a single representative in $C_T = \Pi_C(S - \{x_m\})$. We thus have:

$$|C_S| = |C_T| + |C_*|.$$

Trivially, every concept in C_S appears in C_T, and if it appears twice it is also counted in C_*.

We now bound C_T and C_* separately. The bound for C_T, from the induction hypothesis, is

$$|C_T| \leq J(m - 1, d).$$

We claim that the bound for C_* is,

$$|C_*| \leq J(m - 1, d - 1).$$

Note that if C_* shatters a set $\{x_1, ..., x_i\}$ then C shatters the set $\{x_1, ..., x_i, x_m\}$, since each function can be completed in two different ways. By definition of C_*, for every assignment of $x_1, ..., x_i$ there exist a pair of concepts: $c_0, c_1 \in C$ that are consistent with $c_1(x_m) = 1$ and $c_0(x_m) = 0$. Hence, if C_* shatters a set of size i, then C shatters a set of size $i + 1$. Since we assume $VCdim(C) = d$, then $VCdim(C_*) \leq d - 1$.
Hence, from the induction hypothesis:
$|C_*| \leq J(m - 1, d - 1)$, and

$$|C_S| = |C_T| + |C_*| \leq J(m - 1, d) + J(m - 1, d - 1) = J(m, d),$$

which concludes the proof of the Claim. □

We now have that $|F| \leq J(2m, d)$. Let's explore the function J. Recall that

$$J(m,d) = \sum_{i=0}^{d} \left(\begin{array}{c} m \\ i \end{array} \right).$$

This function has two behaviors:

$$J(m,d) = \sum_{i=0}^{d} \left(\begin{array}{c} m \\ i \end{array} \right) = \left\{ \begin{array}{ll} 2^m & d \geq m \\ 2m^d & d < m. \end{array} \right.$$

That is, the function grows exponentially with m until m reaches d and then it grows exponentially with d. From that we can conclude that the number of functions in the projection can either grow as 2^m or fall to $2m^d$. No intermediate behavior exists.
Hence, back to the bound of m, we now have:

$$\begin{aligned} & m = O(\frac{1}{\epsilon}\log\frac{1}{\delta} + \frac{1}{\epsilon}\log(2m)^d) \\ \Rightarrow \quad & m = O(\frac{1}{\epsilon}\log\frac{1}{\delta} + \frac{d}{\epsilon}\log m) \\ \Rightarrow \quad & m = O(\frac{1}{\epsilon}\log\frac{1}{\delta} + \frac{d}{\epsilon}\log\frac{d}{\epsilon}). \end{aligned}$$

That is, this is the number of samples required to guarantee that in probability $1-\delta$ our error will be smaller than ϵ. As we can see above, we can actually bound the size of the sample with a function of the $VCdim$ alone.
It is also worth noting that the difference between the lower bound and the upper bound we found are relatively small.

8.4 Rademacher Complexity

8.4.1 Motivation

We would like to give tighter bounds for the sample size. Until now our bounds depended on $|\Pi_C(m)|$. How can we estimate $|\Pi_C(m)|$? We can choose a random assignment for S and see if there is a $h \in H$ that corresponds to it. This way, we estimate $|\Pi_C(m)|/2^m$. The problem with this method is that the probability that we "hit" a legal assignment is exponentially small. Another possibility is taking a random assignment and gauging our overfitting. That is, we see how well an $h \in H$ estimates it. For an arbitrary h, we would expect an error of 50%. The best h should give us a better error, say 40%, which would intuitively mean a 10% overfitting.

8.4.2 Problems with VC-dim bounds

The lack of tightness of the VC-dim bounds stems from two weaknesses in our analyses:

1. We use union bound for the samples S_1 and S_2.
2. We choose the worst S for our analysis.

We would like to overcome these two problems. So far we showed that

$$error_D(h) \leq error_S(h) + \alpha\sqrt{\frac{\ln|\Pi_C(2m)|}{m}}.$$

We would like to be give a tighter bound (i.e. replace the rightmost part of the inequality by something smaller).

8.4.3 Rademacher Averages

We now define *Rademacher averages*.

Definition 8.8 *Given a sample $S = \{x_1, x_2, ..., x_m\}$ and a set of functions H, the* empirical Rademacher complexity is

$$R_S(H) = E_\sigma[\max_{h\in H} \frac{1}{m}\sum_{i=1}^{m} \sigma_i h(x_i)],$$

where σ is the random labeling. That is, $\sigma_i \in \{\pm 1\}$, $Pr[\sigma_i = +1] = 1/2$.

Notice that $h(x_i) \in \{\pm 1\}$ as well, and therefore if $\sigma_i h(x_i) = 1$ there is no error ($\sigma_i = h(x_i)$), and if $\sigma_i h(x_i) = -1$ there is an error. When we choose an h that maximizes the sum, we are maximizing the precision (minimizing the number of errors). In other words, we take a random labeling and see how well the "best" h fits it. What we are in fact approximating is, given a random labeling, how good is our hypothesis class.

Definition 8.9 *We define the* Rademacher complexity over a distribution D to be:

$$R_D(H) = E_{S\sim D}[R_S(H)] \ .$$

$R_D(H)$ is the expected value of $R_S(H)$ over a random sample. Note that the complexity depends upon the distribution D and not on the worst sample. The intuition behind the Radamcher average is this: we have a random labeling. What is the best correlation we can get? (Notice that we always get a number between 0 and 1, so $R_D(H) = 1$ means perfect

correlation).
What is the relationship to VC-dim? If C shatters S, what is $R_S(C)$? If C shatters S, each labeling has a $c \in C$ that fits it. Therefore, $R_S(C) = 1$ (for every σ, there is a $c \in C$ such that $\sigma_i = h(x_i)$, therefore the expected value is 1).

Theorem 8.10 *With probability* $1 - \delta$, $\forall h \in H$

$$\begin{aligned} error_D(h) &\leq error_S(h) + R_D(H) + \sqrt{\frac{\ln(2/\delta)}{2m}} \\ &\leq error_S(h) + R_S(H) + 3\sqrt{\frac{\ln(2/\delta)}{2m}}, \end{aligned}$$

where $|S| = 2m$.

Before we prove the theorem, we define McDiarmid's inequality.

8.4.4 McDiarmid's inequality

Theorem 8.11 *Let* $X_1, ..., X_m$ *be independent random variables, and let* $\Phi(X_1, ..., X_m)$ *be a real function such that*

$$\forall i, |\Phi(x_1, ..., x_i, ..., x_m) - \Phi(x_1, ..., x_i', ..., x_m)| \leq c_i,$$

then

$$\forall \epsilon > 0,\ Pr[\Phi(X) > E[\Phi(X)] + \epsilon] \leq exp(-2\epsilon^2 / \sum_{i=1}^{m} c_i^2).$$

In other words, we would like to say that if none of the variables has too much influence on Φ (each variable x_i's influence is less than or equal to c_i), then the probability that a random value of $\Phi(X)$ will be far from its expected value is small.

Corollary 8.12 *If* $\forall i,\ 0 \leq x_i \leq 1$ *and* $\Phi(X) = \frac{1}{m}\sum_{i=1}^{m} x_i$ *, then each* x_i*'s influence is* $\leq 1/m$ *and we get*

$$exp(-2\epsilon^2 / \sum_{i=1}^{m} c_i^2) = exp(-2\epsilon^2 / \frac{m}{m^2}) = e^{-2\epsilon^2 m},$$

which is the Chernoff bound.

Corollary 8.13 *For* $X_i \in [a_i, b_i]$, $\Phi(X) = \frac{1}{m}\sum_{i=1}^{m} x_i$, *we get* $c_i = \frac{b_i - a_i}{m}$ *and*

$$Pr[\Phi(X) > E[\Phi(X)] + \epsilon] \leq exp(-2\epsilon^2 m^2 / \sum_{i=1}^{m} [b_i - a_i]^2),$$

which is Hoeffding's inequality.

8.4.5 Proof of the Theorem

Step 1:

Definition 8.14 *Define* $MAXGAP(S) = \max_{h \in H}[error_D(h) - error_S(h)]$.

We would like to ensure that with probability $1 - \delta$, $MAXGAP(S) \leq \epsilon$.
As the error function is an averaging function, and $|S| = m$, the dependency of $MAXGAP$ on each $x_i \in S$ is at most $1/m$. Therefore we can apply McDiarmid's inequality to $MAXGAP$. Taking $\delta = 2e^{-2\epsilon^2 m}$, we get that with probability $1 - \delta/2$,

$$MAXGAP(S) \leq E_S[MAXGAP(S)] + \sqrt{\frac{ln(2/\delta)}{m}}.$$

If we now show that $R_D(H) \geq E_S(MAXGAP(S))$, this will complete the proof of Theorem 8.10.

Step 2:

To show that $R_D(H) \geq E_S(MAXGAP(S))$, we add another sample S' of size m, where S' is i.i.d. and independednt of S. (S' is referred to as a *ghost sample*). Then

$$error_D(h) = E_{S'}[error_{S'}(h)].$$

Because S is independent of S', we can look at S's error as an average of S'. Let $S = \{x_1, ..., x_m\}$, and $S' = \{x'_1, ..., x'_m\}$.

$$\begin{aligned} E_S[MAXGAP(S)] &= E_S[\max_{h \in H} E_{S'}[error_{S'}(h) - error_S(h)]] \\ &\leq E_{S,S'}[\max_{h \in H}[error_{S'}(h) - error_S(h)]] \\ &= E_{S,S'}[\max_{h \in H} \frac{1}{m} \sum_{i=1}^{m} error_{x'_1}(h) - error_{x_i}(h)]. \end{aligned}$$

($error_x(h)$ is whether h makes a mistake on x. If h makes a mistake, $error_x(h) = 1$, otherwise, $error_x(h) = 0$.)
As before, we first sample $S \cup S'$ and only then split the sample into S and S'. We do this by arbitrarily splitting the sample into m pairs, x_i and x'_i and and then for each pair, deciding which one is in S and which one is in S'. Sampling both sets and then arbitrarily splitting them up yields the same distribution as independently sampling two sets. σ decides which set gets which samples. That is, for each i, if $\sigma_i = -1$, then $x_i \in S$ and $x'_i \in S'$. If $\sigma_i = 1$, then $x_i \in S'$ and $x'_i \in S$.
Continuing with our analysis:

$$E_S[MAXGAP(S)] \leq E_{S',\sigma}[\max_{h \in H} \frac{1}{m} \sum_{i=1}^{m} \sigma_i error_{x'_i}(h)] - E_{S,\sigma}[\min_{h \in H} \frac{1}{m} \sum_{i=1}^{m} \sigma_i error_{x_i}(h)]$$

$$= \frac{2}{m} E_{S,\sigma}[\max_{h \in H} \sum_{i=1}^{m} \sigma_i error_{x_i}(h)]. \tag{8.1}$$

The last equality stems from the symmetry, as $\sigma = \{\pm 1\}$.
We are almost done with our proof. We notice now that in the Rademacher complexity we have $\sum_{i=1}^{m} \sigma_i h(x_i)$. This can be written as $< \sigma, h >$. In our last inequality we have

$$< \sigma, error(h) > \; = \; \sum_{i=1}^{m} \sigma_i error_{x_i}(h) = \sum_{i=1}^{m} \sigma_i I(h(x_i) \neq c_t(xi)).$$

If, in the Rademacher complexity, we write $< \sigma \cdot c_t, h >$ instead of $< \sigma, h >$, we get the same distribution, because σ creates all the vectors, and if we multiply each coordinate by $c_t(x_i)$, it does not affect the distribution.

$$< \sigma \cdot c_t, h > \; = \; \sum(\sigma_i \cdot c^*(x_i)) \cdot h(x_i) = \sum \sigma_i \cdot (c^*(x_i) \cdot h(x_i)) \; = < \sigma, c^* \cdot h >.$$

And

$$c^*(x_i) \cdot h(x_i) = 1 - 2error_{x_i}(h).$$

That is, if there is an error, we get -1, otherwise we get 1.
Putting it all together, we get

$$R_D(H) = E_\sigma[\max_{h \in H} \sum_{i=1}^{m} \sigma_i (1 - 2error_{x_i}(h)/m)]$$

$$= 2E_{S,\sigma}[\max_{h \in H} \sum_{i=1}^{m} \sigma_i error_{x_i}(h)/m]. \tag{8.2}$$

And so, from equations 8.1 and 8.2, we get :

$$E_S[MAXGAP(S)] \leq R_D(H),$$

which completes the proof.
To sum up,

$$error_D(h) \leq error_S(h) + R_D(H) + \sqrt{\tfrac{ln(2/\delta)}{m}},$$

Which gives us

$$R_D(H) \leq R_S(H) + R_D(H) + \sqrt{\tfrac{4ln(2/\delta)}{m}},$$

Using McDiarmid's inequality for $c_i = \frac{2}{m}$ and error ϵ, noting that $exp(-\frac{1}{2}\epsilon^2 m) = exp(\frac{-2\epsilon^2}{m(4/m^2)})$. when $\epsilon = \sqrt{\frac{4ln(2/\delta)}{m}}$ we get $\delta = e^{-2\epsilon^2 m}$